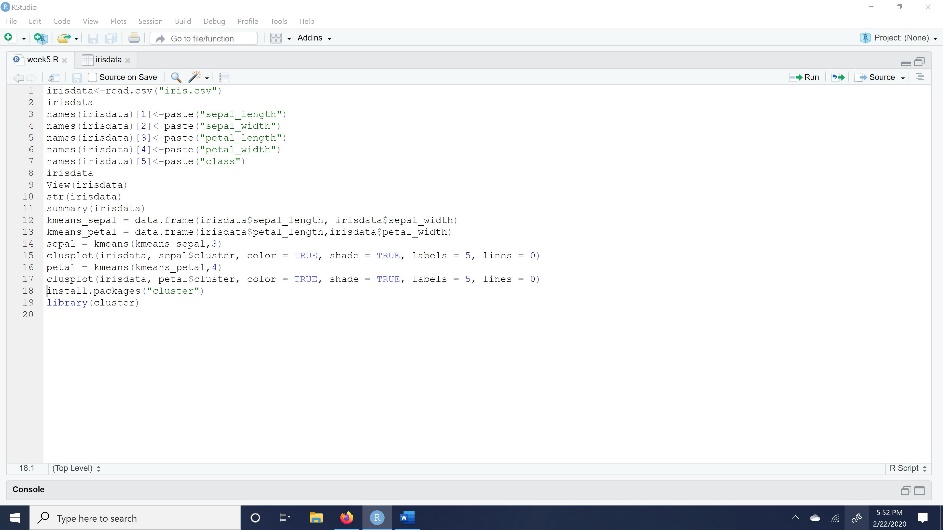
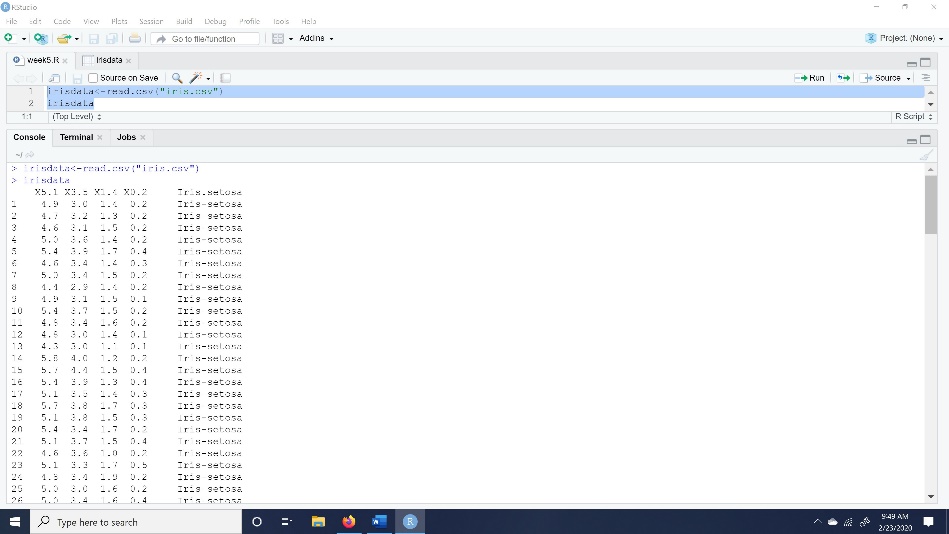
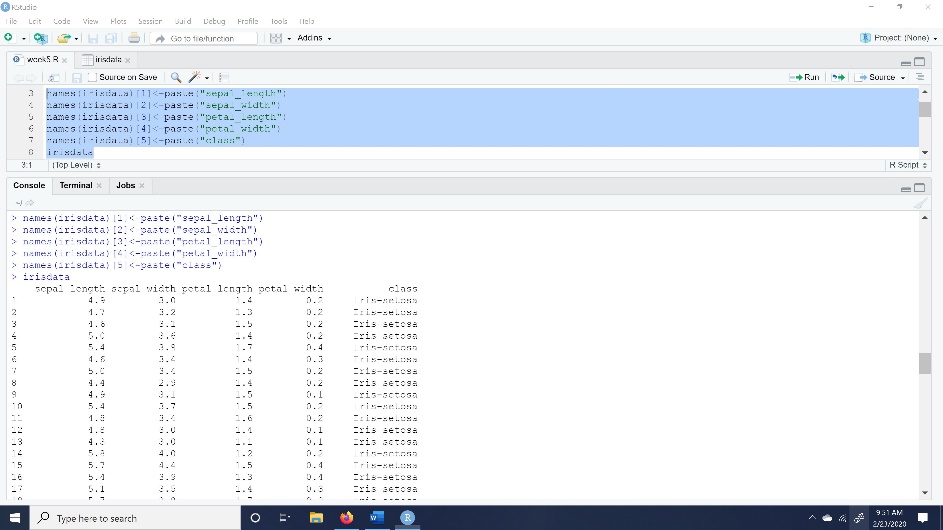
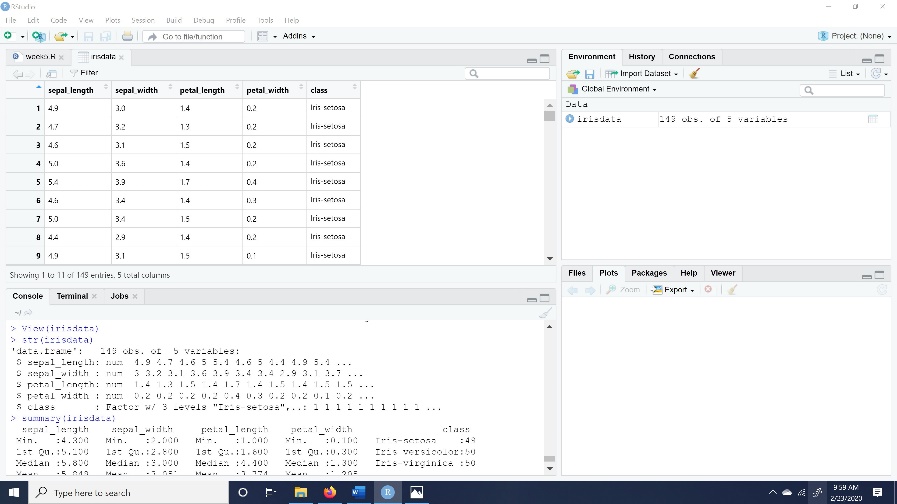
**R Commands:**

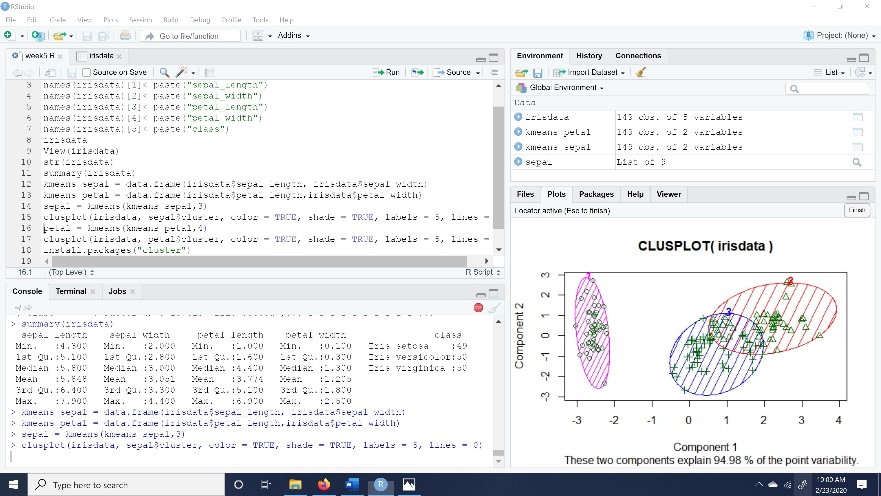


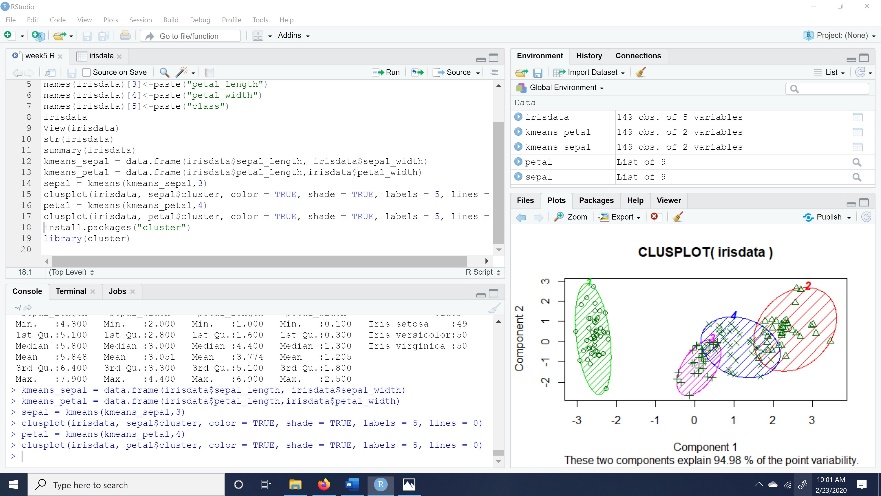
**Outcomes:**



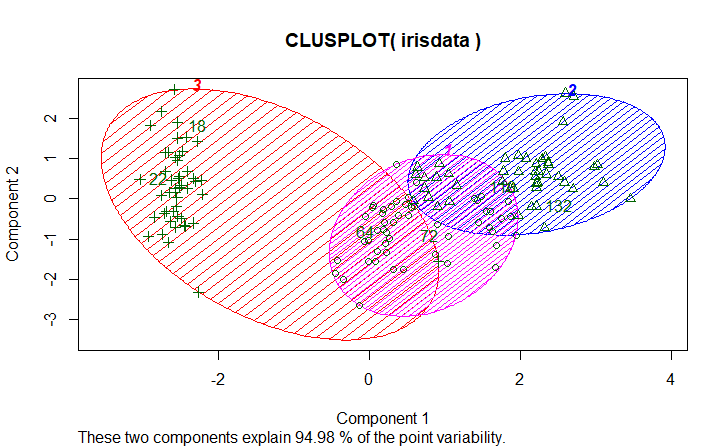


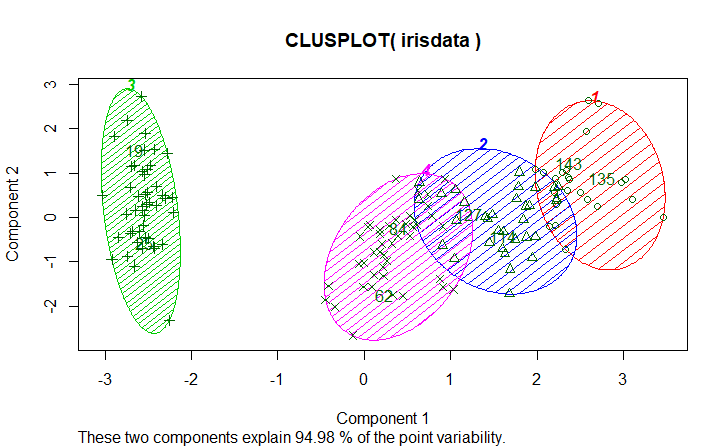






**Plots of both k-means cluster models:**





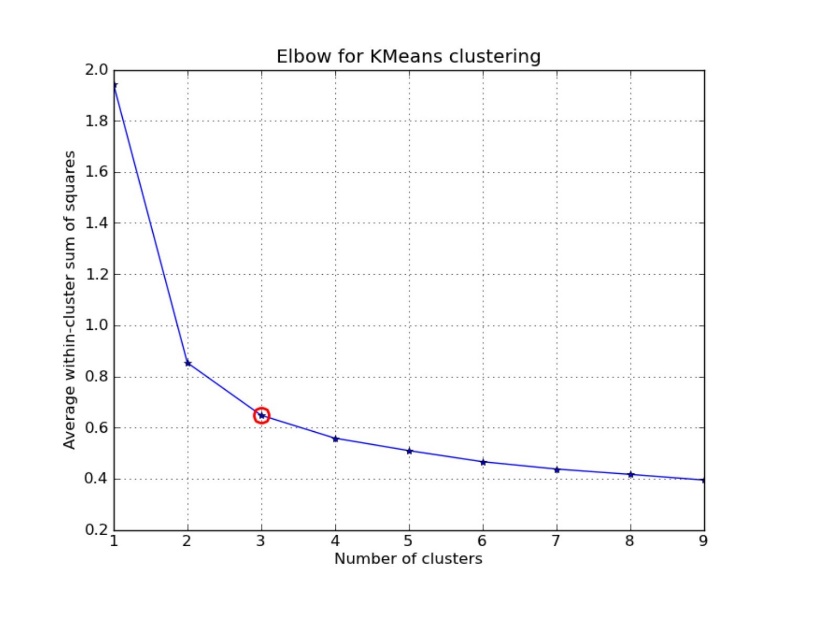
Inter – class and intra – class cluster similarity is a crucial part in clustering. Inter – class cluster show the distance between data point with cluster center, intra – class cluster show the distance between the data point of one cluster with the other data point in another cluster. To produce high quality clusters, we need to have high intra-class similarity and low inter – class similarity.

Reference:

<https://www.cs.cmu.edu/~epxing/Class/10701-08s/Lecture/lecture15-annotated.pdf>

For choosing cluster centroid values we have two methods to decide how many cluster centroids can we take for a dataset. They are Elbow method and silhouette analysis.

**Elbow method** in this method we will try different values for k and plot the average distance of data point from their respective centroid as show in the image below. Let’s look how the value of k varies from 1 to 2 drastically lowers the average distance of each point to its respective centroid. The drawback is we still depends on a manual decision to where the elbow lies. So, this method does not valid for non-clear distinction of an elbow( for instance, when the curve is very smooth).



**Silhouette Analysis** in this automated approach we would be building a collection of k means clustering models with a range of k values and then we examine each model to determine the optimal number of clusters. A silhouette coefficient is calculated for observation which is then averaged to determine the silhouette score. This coefficient combines the average within cluster distance with average nearest-cluster distance to assign a value between -1 and 1. Value that is below 0 indicates that the observation might be a wrong cluster whereas the value that is nearer to 1 indicated that the observation is apt for cluster and this clearly separated from other clusters. This coefficient essentially measures how close an observation is to the nearby clusters, where it is desirable to be the maximum distance possible from nearby clusters. We can determine the best number of clusters, k by choosing the model which has the highest silhouette score.

Reference:

<https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>